Molecular Phylogenetics and Evolution 65 (2012) 765–773

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The effect of sampling from subdivided populations on species identification with DNA barcodes using a Bayesian statistical approach

Melanie Lou, G. Brian Golding *

Department of Biology, McMaster University, Hamilton, Ontario, Canada L8S 4K1

ARTICLE INFO

Article history:
Received 26 April 2012
Revised 30 July 2012
Accepted 31 July 2012
Available online 11 August 2012

Keywords:
Geographical sampling
Population subdivision
Incomplete reference database
Isolation by distance
Segregating sites
DNA barcoding

ABSTRACT

Barcoding is an initiative to define a standard fragment of DNA to be used to assign sequences of unknown origin to existing known species whose sequences are recorded in databases. This is a difficult task when species are closely related and individuals of these species might have more than one origin. Using a previously introduced Bayesian statistical tree-less assignment algorithm based on segregating sites, we examine how it functions in the presence of hidden population subdivision with closely related species using simulations. Not surprisingly, adding samples to the database from a greater proportion of the species range leads to a consistently higher number of accurate results. Without such samples, query sequences that originate from outside of the sampled range are easily misinterpreted as coming from other species. However, we show that even the addition of a single sample from a different subpopulation is sufficient to greatly increase the probability of placement of unknown queries into the correct species group. This study highlights the importance of broad sampling, even with five reference samples per species, in the creation of a reference database.

1. Introduction

DNA barcoding has become a popular method for species identification and delimitation due to advances in the speed and cost of sequencing and the difficulty in delineating unknown specimens using traditional criteria. In addition to proper biodiversity assessment, barcoding has important implications in various areas such as: effective monitoring of invasive and pest species, identifying disease vectors and protecting consumers from market substitutions (Ball and Armstrong, 2006; Lowenstein et al., 2010; Wong et al., 2011).

Since its initial introduction (Hebert et al., 2003), the initiative has evolved from using a distance-based threshold to using a variety of different evolutionary signals to resolve species boundaries (Abdo and Golding, 2007; Davis and Nixon, 1992; Hebert et al., 2003, 2004b; Lou and Golding, 2010; Munch et al., 2008; Ratnasingham and Hebert, 2007; Sarkar et al., 2008). Furthermore, an increasing number of studies advocate the use of traditional lines of evidence (whether behavioral, ecological, geographical, morphological or reproductive) in combination with sequence data to provide further support by showing a correspondence between the two. This combined use of barcoding data with other forms of information has resulted in several well-supported studies that may not have been as reliable if the delimitations had relied solely on sequence data (DeSalle et al., 2005; Siddall and Budinoff, 2005).

The use of additional information may become essential when problems occur from reference sequence data with low information content. One of the benefits of using a mitochondrial marker is that we expect it to better reflect species boundaries because the expected time to obtaining clear, distinct species groups (i.e. reciprocal monophyly) is short because of its small effective population size (Neigel and Avise, 1986). However, population subdivision with limited gene flow can increase the time to coalescence and, consequently, the time required to achieve reciprocal monophyly (Hudson and Coyne, 2002; Wakeley, 2000). With a lengthened time to reciprocal monophyly, lineages between less closely related species may coalesce before lineages within a species (a phenomenon known as incomplete lineage sorting; (Neigel and Avise, 1986)), thus blurring species boundaries and impeding accurate species diagnoses and delimitations. This is particularly problematic with recently diverged species, a group already prone to incomplete lineage sorting, where the effects of subdivision and migration are more pronounced (Wakeley, 2000). Wong et al. (2011) suggested that some incorrect delimitations reflected the failure to consider the geographic divergence of catfish. Similarly, Papadopoulou et al. (2008) have shown that different rates of gene flow greatly affect divergences and is one of the reasons that can cause DNA barcoding failures. While the effects of subdivision are explored here, it should be noted that mtDNA is not a perfect marker and may occasionally also show non-neutral evolution,

* Corresponding author. Fax: +1 905 522 6066.
E-mail addresses: mlou@evol.mcmaster.ca (M. Lou), Golding@McMaster.CA (G.B. Golding).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.07.033

non-clonal inheritance and variation in mutation rates (Galtier et al., 2009).

One way to acknowledge hidden population subdivision is to sample sequences from across a broad geographical range. Any within-species variation is likely to be widely distributed among several geographical localities or demes and sampling this variation is crucial to being able to correctly calculate the probability of origin and distinguish between close sister species. Many barcoding difficulties may, in part, be due to the failure to choose an appropriate sampling scheme (Meier et al., 2006; Meyer and Paulay, 2005; Wiemers and Fiedler, 2007; Wong et al., 2011). Inherent within-species variation may be spread across local, geographical populations of individuals of one species and, by employing a broad sampling scheme, the addition of these dispersed individuals should aid barcoding identification, provided that the sampled sequences sufficiently reflect the variation within the species.

The effect of sampling on identification and delimitation has been investigated in distance, tree, and general mixed Yule-coalescent (GYMC) methods (Bergsten et al., 2012; Hendrich et al., 2010; Meyer and Paulay, 2005; Monaghan et al., 2009; Ross et al., 2008; Virgilio et al., 2010; Zhang et al., 2010). Further complexities have also been taken in account, for example, Bergsten et al. (2012) has investigated sampling strategies ranging from a local to global scale and Zhang et al. (2010) has investigated sampling from two different models of population structure: a linear stepping-stone and an equilibrium island model with unequal sample sizes in three subpopulations. However, no study, to date, has been conducted using a Bayesian statistical method capable of providing an assessment of identification confidence. While Bergsten et al. (2012) used a threshold value to calculate the proportion of ambiguous assignments (i.e. the number of queries assigned to more than one reference species) as a measure of method uncertainty, it is not as statistically accurate as a Bayesian method where the probability of assignment describes the assignment to a particular species given that it could also assign to other species possibilities. Setting the classification within a statistical framework to generate posterior probabilities is preferred since difficulties in the classification of sequences from very recently diverged sibling species are expected via any methodology. We previously introduced the segregating sites algorithm, a fast, Bayesian tree-less method that is able to calculate the probability that the sequence might originate (`PrOr`) from any one of the candidate species (Lou and Golding, 2010). Due to its speed and the large body of theory behind it, the segregating sites algorithm is further explored in this paper to investigate the efficacy of this algorithm when species have recently diverged and exist in subdivided groups.

Here the identification performance of DNA barcodes with broader samples is analyzed using our Bayesian statistical method, the segregating sites algorithm (Lou and Golding, 2010), in a population structure model based on isolation by distance. To investigate the efficacy of barcoding in species with population substructure, we simulated sequences based on three parameters: a sampling scheme of reference sequences (to represent differences in the number and location of dispersed samples among demes), rates of migration between demes and times to divergence between species. For various combinations of these parameters, we examined the probability that a query sequence originates from each species as calculated by the segregating sites algorithm. As an application, the same testing procedure was carried out with cytochrome c oxidase subunit 1 (*CO1*) sequences of the genus *Grammia* (Lepidoptera: Noctuidae). The tiger moth species of this genus provides a good case study where classical morpho- and ecological traits do not agree with species groupings based on mitochondrial DNA (mtDNA). As 54% of the sampled species share haplotypes with at least one other species, under the barcoding gap criterion that no overlap between intra- and interspecies divergences be present, this would result in incorrect diagnoses for 32% of the species (Schmidt and Sperling, 2008). Both our simulated results and the empirical findings show that including at least one dispersed sample can aid sequence identification, even with recently diverged species and that including more dispersed samples further improves these results.

Our results highlight the importance of considering population subdivision and gene flow to the barcoding workflow, particularly for species known to have wide distribution ranges, and to sample broadly whenever possible to ensure that representative samples that contribute to describing the species boundary are included. Minimally, the results show that a single extra sample from another locality goes a long way to ensure accuracy.

2. Methods and data

2.1. Population spatial substructure

The simulation is based on an isolation by distance population model where every individual is restricted in its local movement to neighboring demes (two-dimensional movement within a $d \times d$ square lattice where $d \times d$ represents the number of demes). Therefore, individuals are much more closely related to nearby individuals than to distant individuals. Let $\theta = 4N_e\mu$ be the population mutation rate (μ is the mutation rate per locus per generation), $M = 4N_em$ be the symmetric migration rate between demes (m is the proportion of the population that migrates between two demes per generation) and N_e is the effective population size. All demes are assumed to be of constant and equal size. The taxonomy of the reference sequence data is assumed to be correct.

2.2. Coalescent model with population substructure

Let each species exist within its own lattice and let each deme within the lattice contain any number of sampled sequences from the species. In generating a coalescent history of the lineages, the occurrence of a coalescent or migration event depends on where the lineages exist on the lattice. The probability of a coalescent event is more likely if many lineages are found within the same deme; otherwise a migration event is more likely. Coalescent theory with a consideration for population structure is well developed. The times until a coalescent or migration event are exponentially distributed with means:

$$I_{coal} = d\sum_{i=1}^{d}\frac{k_i(k_i-1)}{2}$$

and

$$I_{migr} = \frac{Mdk}{2}$$

respectively (where k_i represents the number of lineages in deme i and $k = \sum_{i=1}^{d} k_i$ is the total number of lineages in all demes; Hein et al. (2005)).

The sum of the above two, $I_{coal} + I_{migr}$, represents the total rate until the occurrence of an event and the probability that the next event is a coalescent event or migration event is:

$$\frac{I_{coal}}{I_{migr}+I_{coal}} = \frac{\sum_{i=1}^{d}k_i^2 - k}{k(M-1)+\sum_{i=1}^{d}k_i^2}$$

and

$$\frac{I_{migr}}{I_{migr}+I_{coal}} = \frac{kM}{k(M-1)+\sum_{i=1}^{d}k_i^2}$$

respectively. For further details, refer to Hein et al. (2005).

2.3. Simulation

We simulated a multi-species coalescent (Degnan and Rosenberg, 2009), based on a total of 10 species. Each species has five sampled sequences. Five is the recommended minimum by the Consortium for the Barcode of Life (CBOL) (Hajibabaei et al., 2007) and via simulation study (Ross et al., 2008). The first species has one additional sampled sequence, which is used as the unknown query sequence. Each of the remaining nine species are progressively more and more distant from the first species (in a pectinate or asymmetric pattern). Other patterns were simulated with qualitatively similar results. These simulations permit incomplete lineage sorting but do not address introgression. Two lineages can coalesce only if their sequences exist within the same deme. Going back in time, the lineages will coalesce at a rate determined by the population size and migration rates. At a predetermined time, T, speciation is assumed to occur. At this time, lineages of either species, whether coalesced to a single ancestor or not, are randomly placed on this new lattice and thereafter treated as a single species. This process is repeated until a full coalescent history of all 10 species is obtained.

Once the full coalescent is constructed random substitutions are placed on the branches of the coalescent, according to the rate θ, and the resulting sequence data at the leaves are taken as the simulated data.

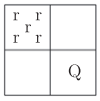
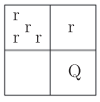
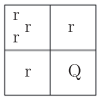
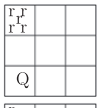
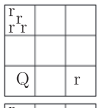
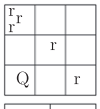
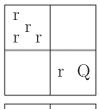
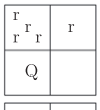
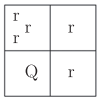
Given 51 simulated reference sequence data, the query sequence is removed from the reference data set and it, along with the remaining simulated sequences, are tested by the segregating sites algorithm (Lou and Golding, 2010) to determine the probabilities of origin (`PrOr`) for the query from each of the 10 species. We have previously shown that the segregating sites algorithm can reliably assign unknown specimens even in the absence of a barcoding gap (a separation between intra- and interspecific variation).

We hypothesize that the probability that the query sequence originated from the first species should be greater when at least one or more dispersed sequences are included in the analysis. A sequence from the correct species but located in a spatially distinct deme adds important intraspecific variation that would not be obtained if all the reference samples originate from a single deme. At a minimum, the number of simulations where the `PrOr` is highest for the first species should be at least equal to the number where the first species is monophyletic. This should represent a minimum expectation.

2.4. Simulated data

The sampling scheme of reference sequences on the lattice, the number of demes, time to coalescence, and rates of migration are allowed to vary. We set the DNA sequence length equal to 600 bp, θ to 2.0, and modelled 10 species, each represented by five lineages. The sequence length chosen is approximately the length of the 648-bp barcoding region (Hebert et al., 2004b) and the level of sequence variation (θ) was chosen to be sufficient so that it

Table 1
Lattice sampling schemes analyzed. Each `r` represents a reference sequence belonging to the species from which the query sequence, `Q`, originates. By default, $d \times d = 4$ while the suffix 'L' sets $d \times d = 9$ (see `all_L`, `1other_L`, `2other_L`).

Sampling scheme	Lattice layout	Description
all		`all` reference sequences from one, base, sampling region and query, `Q`, in region furthest from the base
1other		One dispersed reference sequence adjacent to the base region
2other		Two dispersed reference sequences in independent regions, adjacent to the base region
all_L		`all` reference sequences from base region; $d \times d = 9$
1other_L		One dispersed reference sequence from a region furthest from the base; $d \times d = 9$
2other_L		Two dispersed reference sequences in independent regions: one is in the center deme of the lattice and the other is in from a region furthest from the base; $d \times d = 9$
withQ		One dispersed reference sequence in the same region as the query
Qcloser_1other		One dispersed reference and query sequence in independent regions, adjacent to the base region
Qcloser_2other		Two dispersed reference and query sequences in independent regions, adjacent to the base region; the query is closer to the base region

mimics a marker like *CO1* that is able to discriminate at the species level and yet remain relatively conserved given its indispensable role in energy production (Capaldi, 1990). Simulations show that the number of alleles sampled per locus does not have a significant effect on the time to coalescents that exhibit reciprocal monophyly (Hudson and Coyne, 2002; Knowles and Carstens, 2007).

Table 1 shows the various sampling schemes for the reference species; all configurations are placed in a lattice containing 4 demes (2×2) or 9 demes (3×3). The scheme `all` represents a sampling situation where all the reference sequences are from one deme or base region. The schemes `1other` and `2other` represent situations where one or two dispersed samples, respectively, are included in the reference species. We were also interested in the effect of a larger lattice or sampling area (`all_L`, `1other_L` and `2other_L` where $d \times d = 9$). We also investigated the effect of query placement, relative to the base region (`Qcloser_1other` and `Qcloser_2other`) and, lastly, a configuration where a reference and query sample originate from the same deme (`withQ`).

The time to speciation (backward-in-time), T, was set to 10 and 3 (scaled in units of $2N_e d{*}d$ generations). The symmetric rate of migration, M, ranged from 0.1 to 1000 (M up to 10 shown here), to model different rates of movement among demes within a lattice. When the time to speciation is long and the migration rate is high, the lineages within each species should coalesce with each other first and the level of variation within a species should be less than between species; this represents the ideal situation where each population is a distinct and monophyletic species (Avise, 1989) and we expect most of these simulations to have `PrOr` largest for the first species. When the time to speciation is short or the migration rate is low, there will be more incomplete lineage sorting and this would result in a lower proportion of the simulations where the `PrOr` is largest for the first species. Each combination of parameters are based on 10,000 simulation runs.

Since these simulations are conducted within a statistical framework, we have the advantage of not only identifying correct assignments but also those that occur with high confidence. Thus, to be conservative, we additionally considered analyses of simulations where the `PrOr` is $\geqslant$80%. Due to difficulties with obtaining simulations that satisfied this criterion, these results are based on 100 simulation runs. The difficulty arises because conspecific lineages will take a long time to coalesce if they are spread among many demes that seldom migrate when the migration rate is low, thereby increasing the chance of paraphyletic coalescents (Wakeley, 2000).

2.5. Empirical data: Grammia

Species of the *Grammia* genus have a large geographic range, exhibit interspecific hybridization and incomplete lineage sorting, making them an ideal data set to explore the use of dispersed samples on assignment fidelity. Of several *Grammia* species for which sequence information is available, we've chosen *Grammia nevadensis* as our focal species because of the paraphyly of its lineages with those from most of the species in the Western clade. It has been widely sampled from 16 locations spanning several provinces of Canada and northwestern US states (Schmidt, 2009; Schmidt and Sperling, 2008).

All 225 *Grammia* sequences from 33 species (Schmidt, 2009; Schmidt and Sperling, 2008) were downloaded from NCBI. Species represented by at least 5 individuals were kept for further analysis. This criterion limited our reference data set to 179 *Grammia* sequences from 13 species (Table 2). For sampling scheme `all`, *G. nevadensis* contained sequences only from British Columbia, and the query was chosen to be from Utah. Dispersed sequences for schemes `1other-5other` are sampled from two provinces in Canada (Alberta, Saskatchewan) and three states in the US (Washington, Oregon, Colorado). The inclusion of one or more dispersed sequence(s) was compensated by a reduction of sequences from British Columbia to maintain a total of five reference sequences for the species.

Table 2
Summary of *CO1* data for 12 `Grammia` species (Schmidt, 2009; Schmidt and Sperling, 2008) and for `Holarctia obliterata` which served as an outgroup.

Species	Monophyletic	Sequences
`Grammia arge`	yes	5
`Grammia celia`	no	5
`Grammia figurata`	no	11
`Grammia nevadensis`	no	18
`Grammia ornata`	no	9
`Grammia parthenice`	no	13
`Grammia phyllira`	yes	7
`Grammia quenseli`	no	10
`Grammia virgo`	no	9
`Grammia virguncula`	no	37
`Grammia williamsii`	no	44
`Grammia williamsii tooele`	no	6
`Holarctia obliterata`	yes	5
Total	–	179

3. Results

3.1. Simulation

Using the segregating sites algorithm, an assignment is considered correct when the `PrOr` is highest for the first species. A multispecies coalescent consisting of distinct and monophyletic species should possess sufficient divergence within and among species to permit the correct assignment of the query to the first species. So we expect a higher proportion of correct assignments when a monophyletic coalescent is recovered for the first species relative to a coalescent that is paraphyletic and includes sequences from other species (i.e. when the time to speciation is short and when the migration rate is low). In other words, the proportion of correct assignments should be at least equal to the proportion of monophyletic trees for the first species.

3.1.1. Effect of population subdivision

Here we focus on the results of simulations where the sampling is largely restricted to one deme (`all`) and the species boundaries are not yet clearly distinct ($T = 3.0$). When the migration rate is low ($M = 0.1$; Fig. 1A), the proportion of correct assignments, 35%, is less than the proportion of simulations with the first species monophyletic, 39%. This indicates that at this level of divergence and migration the assignment of an unknown query sequence is difficult and/or misleading when the reference sequences of the species are sampled from just one location. In a larger sampling area, the distance between the base region and the query is larger and we would expect a decrease in the proportion of correct assignments reflecting the lengthened time required for coalescence. While the result is approximately the same (38% correct assignments vs 42% monophyletic coalescents; Fig. 1D), additional simulations with an even larger lattice ($d \times d = 16$) confirmed the prediction (data not shown).

When one dispersed sequence was included in the reference dataset (`1other`, `1other_L` and `Qcloser_1other`; Table 1), the number of simulations where `PrOr` was largest for the first species increased whether sampled in a small (44% vs 26%; Fig. 1B) or large sampling area (39% vs 14%; Fig. 1E) or when the query is sampled closer to the base region (56% vs 24%; Fig. 1H).

When two dispersed sequences were included in the reference dataset (`2other`, `2other_L` and `Qcloser_2other`; Table 1), the

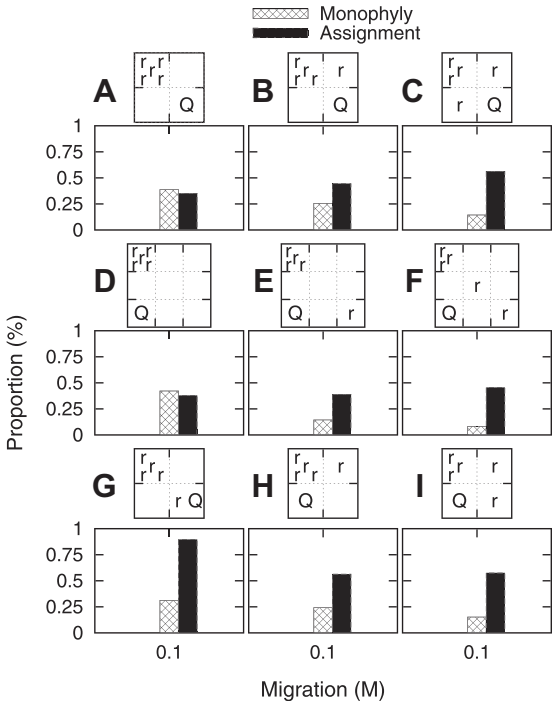


Fig. 1. Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T = 3.0$ and $M = 0.1$. Each subfigure, A-I, has a specific placement of reference (r) and query (Q) sequences for the correct species (Table 1). Monophyly represents the proportion of monophyletic coalescents for the first species (double-hatched bars). Assignment represents the proportion of correct assignments where the query assigned to the first species (solid bars).

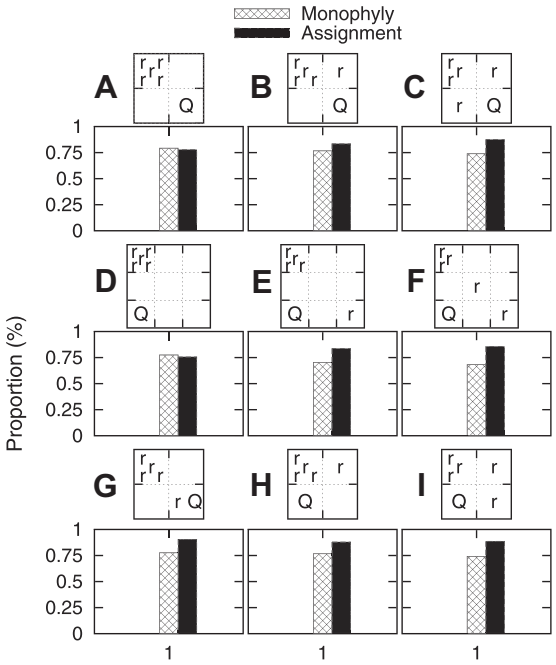


Fig. 2. Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T = 3.0$ and $M = 1$. See Fig. 1 for simulation and legend details.

proportion of simulations with PrOr largest for the first species increased further: in a small (56% vs 14%; Fig. 1C) or large sampling area (46% vs 8%; Fig. 1F) and when the query is closer to the base region (57% vs 15%; Fig. 1I).

The sampling scenario that returned the highest proportion of correct assignments (90% vs 31%) is when the query and a reference sample are sampled from the same deme (Fig. 1G).

3.1.2. Effect of migration

To examine the effect of migration, simulations were repeated with 10-fold increases in the rate of migration. A high migration rate allows lineages to move with greater ease among the demes of a lattice. Consequently, conspecific lineages will coalesce sooner and in turn increase the chances of monophyly.

When the rate of migration was $\geqslant 1$, close to 100% of the simulations were monophyletic for the first species and the PrOr was, correctly, largest for the first species (Figs. 2 and 3). If the rates of migration are large and there is sufficient variation to distinguish species then employing a comprehensive sampling scheme is not necessary.

3.1.3. Conservative assessments

When we restrict our analyses to simulations where the PrOr is strongly supported (PrOr $\geqslant$ 80%), there is a large increase in the number of correct assignments when discrete species are considered (from 46% (A) to 83% (C), Fig. 4 (L)ower; $T = 10$) but the effect disappears for newly divergent species (Fig. 4 (L)ower; $T = 3$).

Sampling schemes with two dispersed sequences often did not return any simulation runs where the PrOr was larger than 80% (Fig. 4C, F, and I of (L)ower; $T = 3$). When the migration rate is low and the time to speciation is short, any additional variation at some point cannot compensate for the increased levels of paraphyly. However, for every sampling scheme, the number of simulations where the PrOr is largest for the first species increase relative to the amount of monophyly. It is simply that the degree of certainty for these has been reduced.

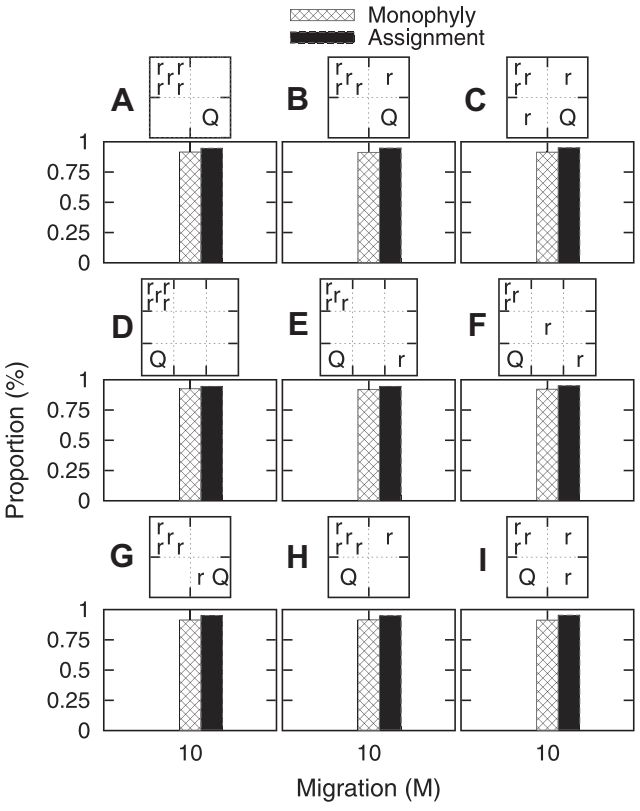


Fig. 3. Inclusion of dispersed samples aids correct identification with recently diverged species. Each histogram is based on 10,000 simulations when $T = 3.0$ and $M = 10$. See Fig. 1 for simulation and legend details.

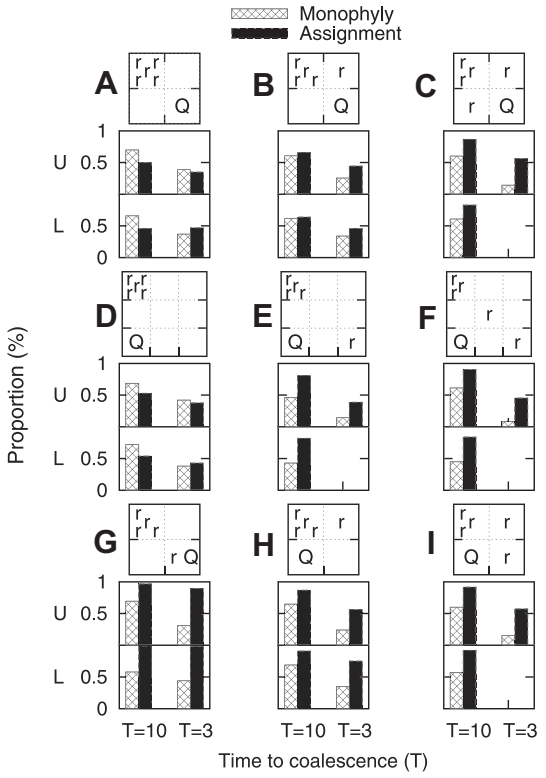


Fig. 4. Simulations based on $M = 0.1$ and T is scaled in units of $2N_e d * d$ generations. The upper row of histograms (U) is based on 10,000 simulations and the lower row of histograms (L) is based on 100 simulations where the `PrOr` is ⩾80%. Monophyly represents the proportion of monophyletic coalescents for the first species (double-hatched bars). Assignment represents the proportion of correct assignments where the query assigned to the first species (solid bars). The inclusion of dispersed samples aids correct identification with high confidence. This is not confirmed for recently diverged species (C, E, F, I) because of a lack of high confidence assignments due to a higher level of paraphyly.

3.1.4. Performance

We wanted to investigate how the probability changes as the composition of the first species is changed to more accurately reflect the variation of species with a wide distribution range when both gene flow and the time to speciation are low ($M = 0.1$ and $T = 3.0$ respectively). To do this, we began with a simulation in which all sequences from the first species are restricted to one deme (`all`; Table 1) and then repetitively change the sequence composition to increasingly reflect a species with a wider distribution (that is, all the sequences are randomly dispersed on the lattice, each individual in its own deme). Each simulation was repeated 10,000 times. Performance is measured as the ratio of the number of simulations observed where `PrOr` is largest to the first species relative to the number where the first species is monophyletic.

As expected, as more dispersed sequences are included, there is a decrease in monophyletic coalescents with a corresponding increase in the number where `PrOr` is largest for the first species. This strongly supports the use of dispersed sequences to form the reference datasets. However, when the correct species is entirely composed of dispersed sequences, the performance decreases from 15 to 13 correct assignments/monophyletic tree (Fig. 5, number of dispersed samples = 5) suggesting that the number of correct assignments returned cannot be expected to do much better when the correct species consists entirely dispersed sequences because of the greater level of paraphyly.

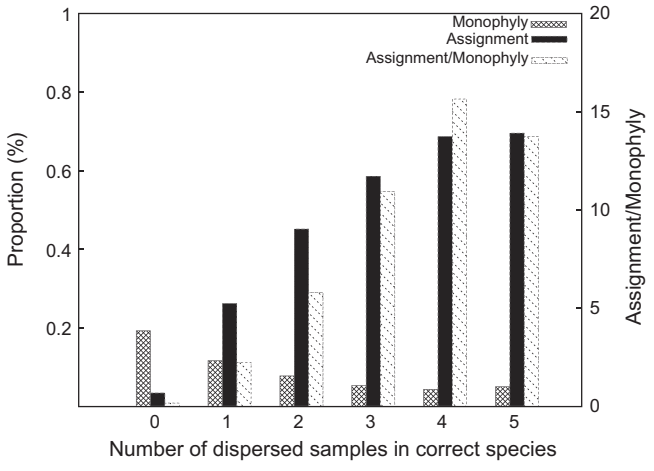


Fig. 5. Increasing performance in assignment when the correct species is composed of more dispersed sequences. Each histogram is based on 10,000 simulations when $M = 0.1$ and $T = 3.0$. When the correct species is entirely composed of dispersed sequences, performance decreases because there is a greater level of paraphyly.

3.2. Grammia (Tiger moth) example

Our analysis of all possible assignments concerning the composition of *G. nevadensis* (Fig. 6 show that the probability of correctly assigning the query increases when at least two dispersed sequences are included (Table 3, columns '% Max' and 'Max P (CA)').

Low probabilities are largely attributed to the extensive paraphyly among western *Grammia* species and, to a lesser extent, the nature of the segregating sites algorithm which calculates high probabilities of assignment to distantly related taxa if they have extensive sequence variation. As expected, the Utah *G. nevadensis* query sequence had high `PrOr` to species found in the Western *Grammia* haplogroup. However, the query consistently had the highest `PrOr` to *Grammia williamsii*. There is extensive sequence variation among the 50 *G. williamsi* specimens that broadly span the US, with some sequences in both Western and Eastern haplotype clades (Schmidt, 2009). Among the 23 haplotypes, some are unique to subspecies *G. williamsii tooele*, some share haplotypes

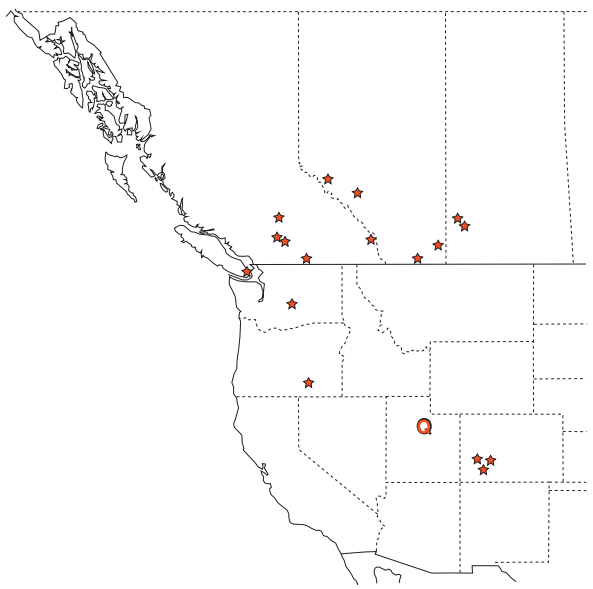


Fig. 6. Geographical locations of *G. nevadensis* samples (stars) and query (Q). Samples are from (with number of sequences in parentheses) British Columbia (6), Alberta (4), Saskatchewan (2), Washington (1), Oregon (1), Colorado (3).

Table 3
Including dispersed sequences for the correct species increases the number of correct assignments. See footnote for details.

Sampling scheme	Total	P (CA)	% Max	Max P (CA)	Mis-assigned to	Avg P
all	1	0.111	0	0.000	*G. williamsii*	0.174
1other	66	0.015	0	0.000	*G. williamsii*	0.192
2other	825	0.142	59	0.178	*G. williamsii*	0.178
3other	3300	0.158	50	0.173	*G. williamsii*	0.168
4other	4950	0.161	52	0.170	*G. williamsii*	0.166
5other	2772	0.162	63	0.168	*G. williamsii*	0.166

Values are based on assignment among 10 species.
In all, the composition of the correct species, *G. nevadensis*, contains 6 samples strictly from British Columbia. The composition of *G. nevadensis* is modified to contain one or more dispersed samples (1other–5other). The dispersed samples are sampled from two provinces in Canada (Alberta, Saskatchewan) and three states in the US (Washington, Oregon, Colorado). The query is from Utah. See Fig. 6 for the geographical locations of the sampled sequences and the query.
Total indicates the total number of possible combinations for assignment. Correct assignment (CA) indicates assignments to the first species (*G. nevadensis*). P (CA) is the average PrOr from *G. nevadensis*. % Max is the proportion of assignments when P (CA) is the largest for *G. nevadensis*. Max P (CA) is the average P (CA) when it is the largest for *G. nevadensis*. If the query is incorrectly assigned, the incorrect species (Mis-assigned to) and the average PrOr from this species (Avg P) is given.

Table 4
For all assignments, correct or incorrect, the statistical risk of assigning to the correct species is always the lowest. This suggests that the query originates from the correct species. See footnote for details.

Sampling scheme	Total	Risk (CA)	% Min	Min risk (CA)
all	1	0.007	100	0.007
1other	66	0.011	100	0.011
2other	825	0.011	100	0.011
3other	3300	0.012	100	0.012
4other	4950	0.013	100	0.013
5other	2772	0.013	100	0.013

See Table 3 for simulation details.
Total indicates the total number of possible combinations for assignment. Correct assignment (CA) indicates assignment to the first species (*G. nevadensis*). Risk (CA) is the average statistical risk of assignment to *G. nevadensis*. % Min is the proportion of assignments when Risk (CA) is the lowest for *G. nevadensis*. Min risk (CA) is the average Risk (CA) of assignment when it is the lowest for *G. nevadensis*.

with a few Eastern clade species, and some have introgressed with other species (Schmidt, 2009; Schmidt and Sperling, 2008). Because of its hyper-variability, *G. williamsi* acts as a single, morphological species that is capable of generating a coalescent that includes any query. In all cases, however, the statistical risk is always lowest for *G. nevadensis* (Table 4. Minimal statistical risk (a metric included in the segregating sites algorithm) to *G. nevadensis* suggests that the 'loss' of assigning the query to *G. nevadensis*, given that it could assign to other species, is small and that it is the species of origin (see Abdo and Golding, 2007, for further details).

4. Discussion

Barcoding with the mitochondrial *CO1* gene sequence has been successful in many groups of animals (Hebert et al., 2004a) but has proved less successful in some other groups (Meyer and Paulay, 2005; Monaghan et al., 2005). The problem is the lack of correspondence between sequence-delimited groups and taxonomically recognized species. This lack of agreement is attributable to a variety of phenomena such as incomplete lineage sorting (Hudson and Coyne, 2002), allopatric speciation (Coyne and Orr, 2004), gene- and species-tree discordance (Funk and Omland, 2003) and the criteria used to determine species boundaries (Mayr, 1942). Other practical problems include incomplete reference databases with insufficient within-species sampling, which is required for accurate species authentication (Meyer and Paulay, 2005; Siddall and Budinoff, 2005).

When gene flow is high and when species divergence times are large, methods to classify sequences to species groups should be relatively straightforward. However, when gene flow is low and species divergence times are small the ability to correctly classify sequences will then be impaired. In these situations, methods that determine the posterior probability that the sequence originates from each species become critical.

We would expect that the number of simulations with the highest probability of origin (PrOr) to the first species should be at least equal to or larger than the number with a monophyletic relationship among the reference sequences. However, our results suggest that this is not always true. Whenever, significant population subdivision exists and reference sequences have not been collected from different demes, then a query sequence from a different deme will appear sufficiently different from the reference sequences to prevent correct identification (Fig. 1A). On the other hand, adding just a single reference sequence from a divergent deme can reverse this and the PrOr will be higher than the proportion of monophyletic reference species (Fig. 1B). This is a result of the increased estimate of conspecific variation as represented by increased θ values. Continuing to add more samples from divergent demes further improves the relative ratios (Fig. 5).

The tiger moth species of the *Grammia* genus are an example of a group with geographically widespread populations connected by gene flow. Despite extensive non-monophyly among these species, the PrOr from *G. nevadensis* increased when the database contained dispersed samples spanning the geographic locales between the base sampling region (British Columbia) and the origin of the query (Utah) (Fig. 6).

Both the simulation and empirical results suggest that the success of mtDNA barcodes depend on sufficient reference sequences that are representative of the within-species variation and when it is undersampled, from substructured genetic variation (population subdivision) or newly divergent species or both, inaccurate species identifications and delimitations may result. This reflects the requirement from traditional taxonomy to ensure that sufficient variation is sampled in order to determine if characters are taxonomically useful (DeWalt, 2011; Trewick, 2007; Wong et al., 2011). Thus, mtDNA sequence is a valuable tool but only with a comprehensive database consisting of complete conspecific reference sequences, especially from species with wide geographical distributions or that have recently diverged, and our study attests to the need for methods to consider adequate representation of the natural variation within the species.

Furthermore, accurate species delimitations have important implications in the development of proper guidelines and policies used to manage and protect both biodiversity and consumer interests. This includes areas such as, but not limited to, conservation and disease biology and aquaculture.

Our results could be expanded upon to allow the coalescent–speciation transitions to vary in space (e.g. along different branches of the tree; (Monaghan et al., 2009)) and in time (e.g. unsampled lineages in demes that have gone extinct; (Lohse, 2009)). The examination of peripheral populations is of particular importance for recently speciated groups. We also assumed that lineages

migrate in a discrete and symmetric fashion but it would be more accurate to model continuous movement among demes. A recent method by Lemey et al. (2010) uses a continuous spatial diffusion model to identify the ancestral geographical history of a sample of sequences but may not be applicable in our simulations since it is not meant to infer population-based spatial histories (Bloomquist et al., 2010). Although the current model is limited in these respects, it is sufficient to illustrate how broad sampling of within-species divergence is essential for accurate barcoding identifications, how this variation affects identifications, and that even minimal sampling goes a long way.

While it is important to include singletons ([species described by a single sample;]Lim et al., 2011) in the biodiversity inventory, a singleton cannot capture any of the variation or complexity of a species (Ross et al., 2008). This variation is critical for any population genetic method, such as segregating sites algorithm, that describes the conspecific variation via a summary statistic (θ) and this calculation requires multiple samples. For this reason, singletons are excluded from the reference data set used here. However, despite their exclusion, if extraneous information is used to estimate this variation then a Bayesian method such as the segregating sites algorithm should be able to identify queries that originate from singletons in the reference database. One such source of extraneous information might be to assume that the singleton species has a level of variation (θ) equal to that of sibling species. At the other end of the sampling size spectrum, Bergsten et al. (2012) recommended a minimum of 20 samples per species for any sampling strategy. However, the authors also note that the choice of the identification algorithm will determine acceptable sample sizes, identification performance, and error rates. Furthermore, Zhang et al. (2010) found that a universal sample size is unrealistic for different species and that it ultimately depends on the evolutionary history of the species. By evaluating the segregating sites algorithm via simulations, we assess its general performance across a range of evolutionary scenarios without particular focus on the *CO1* gene and we find that while more samples will provide better results, a large improvement in the number of correct assignments can be achieved with even a single dispersed sample from a total of five samples per species.

5. Conclusions

Using the segregating sites algorithm and a minimum five samples per species, both simulated and *Grammia* (tiger moth) analyses show that ensuring at least one reference sequence is sampled from a different region or deme of a species distribution returns a greater proportion of results that correctly assign an unknown specimen to its species of origin. Our results highlight the importance of broad sampling to improve the information content of reference samples and that a single dispersed sample can greatly improve the identification of sequences to species.

Accession numbers

179 sequences, 13 species:

AF549620-AF549622
EU119425-EU119426
EU119444-EU119449
EU119459-EU119480
EU119482-EU119485
EU119489-EU119496
EU119498-EU119499
EU119510-EU119580
EU119601-EU119602
EU119605-EU119609
EU119612-EU119665

Acknowledgments

We thank Dr. Richard Morton and Dr. Jonathan Dushoff for revisions and helpful comments. This work was supported by Grants from NSERC and GBG was also funded by a Grant from the Government of Canada through Genome Canada and the Ontario Genomics Institute through the Biomonitoring 2.0 Project (OGI-050).

References

Abdo, Z., Golding, G.B., 2007. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol. 56, 44–56.
Avise, J., 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution 43, 1192–1208.
Ball, S.L., Armstrong, K.F., 2006. DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). Can. J. Forest Res. 36, 337–350.
Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G.N., Ribera, I., Nilsson, A.N., Barraclough, T.G., Vogler, A.P., 2012. The effect of geographical scale of sampling on DNA barcoding. Syst. Biol.
Bloomquist, E.W., Lemey, P., Suchard, M.A., 2010. Three roads diverged? Routes to phylogeographic inference. Trends Ecol. Evol. 25, 626–632.
Capaldi, R.A., 1990. Structure and function of cytochrome c oxidase. Annu. Rev Biochem. 59, 569–596.
Coyne, J.A., Orr, H.A., 2004. Speciation. Sinauer & Associates, Sunderland, Massachusetts.
Davis, J., Nixon, K., 1992. Populations, genetic variation, and the delimitation of phylogenetic species. Syst. Biol. 41, 421–435.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
DeSalle, R., Egan, M.G., Siddall, M., 2005. The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos. Trans. Roy. Soc. B. 360, 1905–1916.
DeWalt, R., 2011. DNA barcoding: a taxonomic point of view. J. N. AM. Benthol. Soc. 30, 174–181.
Funk, D., Omland, K., 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34, 397–423.
Galtier, N., Nabholz, B., Glemin, S., Hurst, G., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18, 4541–4550.
Hajibabaei, M., Singer, G.A., Hebert, P.D., Hickey, D.A., 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. Trends Genet. 23, 167–172.
Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004a. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc. Natl. Acad. Sci. USA 101, 14812–14817.
Hebert, P.D.N., Ratnasingham, S., deWaard, J.R., 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Philos. Trans. Roy. Soc. B 270, S96–S99.
Hebert, P.D.N., Stoeckle, M.Y., Zemlak, T.S., Francis, C.M., 2004b. Identification of Birds through DNA barcodes. PLoS Biol. 2, 1657–1663.
Hein, J., Schierup, M.H., Wiuf, C., 2005. Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory. Oxford University Press Inc., Oxford, UK.
Hendrich, L., Pons, J., Ribera, I., Balke, M., 2010. Mitochondrial cox1 sequence data reliably uncover patterns of insect diversity but suffer from high lineage-idiosyncratic error rates. PLoS One 5, e14448.
Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. Evolution 56, 1557–1565.
Knowles, L.L., Carstens, B.C., 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56, 887–895.
Lemey, P., Rambaut, A., Welch, J.J., Suchard, M.A., 2010. Phylogeography takes a relaxed random walk in continuous space and time. Mol. Biol. Evol. 27, 1877–1885.
Lim, G.S., Balke, M., Meier, R., 2011. Determining species boundaries in a world full of rarity: singletons, species delimitation methods. Syst. Biol., 1–5.
Lohse, K., 2009. Can mtDNA barcodes be used to delimit species? A response to Pons et al. (2006).. Syst. Biol. 58, 439–442.
Lou, M., Golding, G.B., 2010. Assigning sequences to species in the absence of large interspecific differences. Mol. Phylogenet. Evol. 56, 187–194.
Lowenstein, J.H., Burger, J., Jeitner, C.W., Amato, G., Kolokotronis, S.O., Gochfeld, M., 2010. DNA barcodes reveal species-specific mercury levels in tuna sushi that pose a health risk to consumers. Biol. Lett. 6, 692–695.
Mayr, E., 1942. Systematics and the Origin of Species. Columbia University Press, New York.
Meier, R., Shiyang, K., Vaidya, G., Ng, P.K., 2006. DNA barcoding and taxonomy in diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. 55, 715–728.
Meyer, C.P., Paulay, G., 2005. DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. 3, 2229–2237.

Monaghan, M.T., Balke, M., Gregory, T.R., Vogler, A.P., 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Philos. Trans. Roy. Soc. B 360, 1925–1933.
Monaghan, M.T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D.J., Lees, D.C., Ranaivosolo, R., Eggleton, P., Barraclough, T.G., Vogler, A.P., 2009. Accelerated species inventory on Madagascar using coalescent-based models of species delineation. Syst. Biol. 58, 298–311.
Munch, K., Boomsma, W., Huelsenbeck, J.P., Willerslev, E., Nielsen, R., 2008. Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst. Biol. 57, 750–757.
Neigel, J.E., Avise, J.C., 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Nevo, E., Karlin, S. (Eds.), Evolutionary Processes and Theory. Academic Press, London, pp. 515–534.
Papadopoulou, A., Bergsten, J., Fujisawa, T., Monaghan, M.T., Barraclough, T.G., Vogler, A.P., 2008. Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. Philos. Trans. Roy. Soc. B 363, 2987–2996.
Ratnasingham, S., Hebert, P.D., 2007. BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol. Ecol. Notes 7, 355–364.
Ross, H.A., Murugan, S., Li, W.L., 2008. Testing the reliability of genetic methods of species identification via simulation. Syst. Biol. 57, 216–230.
Sarkar, I.N., Planet, P.J., Desalle, R., 2008. CAOS software for use in character-based DNA barcoding. Mol. Ecol. Resour. 8, 1256–1259.
Schmidt, B.C., 2009. Taxonomic revision of the genus Grammia Rambur (Lepidoptera: Noctuidae: Arctiinae). Zool. J. Linn. Soc.-Lond. 156, 507–597.
Schmidt, B.C., Sperling, F.A.H., 2008. Widespread decoupling of mtDNA variation and species integrity in Grammia tiger moths (Lepidoptera: Noctuidae). Syst. Entomol. 33, 613–634.
Siddall, M., Budinoff, R., 2005. DNA-barcoding evidence for widespread introductions of a leech from the South American Helobdella triserialis complex. Conserv. Genet. 6, 467–472.
Trewick, S., 2007. DNA barcoding is not enough: mismatch of taxonomy and genealogy in New Zealand grasshoppers (Orthoptera: Arcididae). Cladistics 24, 240–254.
Virgilio, M., Backeljau, T., Nevado, B., De Meyer, M., 2010. Comparative performances of DNA barcoding across insect orders. BMC Bioinform. 11, 206.
Wakeley, J., 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54, 1092–1101.
Wiemers, M., Fiedler, K., 2007. Does the DNA barcoding gap exist? A case study in blue butterflies (Lepidoptera: Lycaenidae). Front. Zool. 4, 1–16.
Wong, L.L., Peatman, E., Lu, J., Kucuktas, H., He, S., Zhou, C., Na-nakorn, U., Liu, Z., 2011. DNA barcoding of catfish: species authentication and phylogenetic assessment. PLoS One 6, e17812.
Zhang, A.B., He, L.J., Crozier, R.H., Muster, C., Zhu, C.D., 2010. Estimating sample sizes for DNA barcoding. Mol. Phylogenet. Evol. 54, 1035–1039.

Glossary

Monophyly: A species that includes an ancestor and all of its descendants.

Paraphyly: A species whose most recent common ancestor includes descendants from another species

Introgression: The transfer of genetic information from one species to another

Incomplete lineage sorting: A discordance between gene genealogies and species phylogenies

Coalescence: A retrospective description of the evolutionary relationship of a set of individuals based on their most recent common ancestor

Bayesian theory: A mathematical theory for determining the probability of an event based on what you know (prior beliefs) with new knowledge (the data itself)

Deme: A local population of individuals of one species